Global Optimization Using Bad Derivatives: Derivative-Free Method for Molecular Energy Minimization

IOAN ANDRICIOAEI, JOHN E. STRAUB*

Department of Chemistry, Boston University, Boston, Massachusetts 02215

Received 24 October 1997; accepted 9 April 1998

ABSTRACT: A general method designed to isolate the global minimum of a multidimensional objective function with multiple minima is presented. The algorithm exploits an integral ''coarse-graining'' transformation of the objective function, U, into a smoothed function with few minima. When the coarse-graining is defined over a cubic neighborhood of length scale ϵ, the *exact* gradient of the smoothed function, $\mathscr{U}_\epsilon$, is a simple three-point finite difference of U. When ϵ is very large, the gradient of $\mathscr{U}_\epsilon$ appears to be a ''bad derivative'' of U. Because the gradient of $\mathscr{U}_\epsilon$ is a simple function of U, minimization on the smoothed surface requires no explicit calculation or differentiation of $\mathscr{U}_\epsilon$. The minimization method is ''derivative-free'' and may be applied to optimization problems involving functions that are not smooth or differentiable. Generalization to functions in high-dimensional space is straightforward. In the context of molecular conformational optimization, the method may be used to minimize the potential energy or, preferably, to maximize the Boltzmann probability function. The algorithm is applied to conformational optimization of a model potential, Lennard–Jones atomic clusters, and a tetrapeptide. © 1998 John Wiley & Sons, Inc. J Comput Chem 19: 1445–1455, 1998

Keywords: conformational optimization; energy minimization; peptides; simulated annealing

* Alfred P. Sloan Research Fellow (1995–1997)

Correspondence to: I. Andricioaei

Contract/grant sponsor: Alfred P. Sloan Foundation and Petroleum Research Fund of the American Cancer Society; contract/grant number: 30601-AC6

Contract/grant sponsor: National Science Foundation; contract/grant number: CHE-9632236

Contract/grant sponsor: National Center for Supercomputing Applications; contract/grant number: CHE-950015N

CCC 0192-8651 / 98 / 131445-11

Introduction

Problems of global optimization are ubiquitous in science. The solution can often be reduced to the the task of identifying the global extremum of a complicated objective function. Examples include the refinement of neural networks, combinatorial optimization problems such as the traveling salesman problem and various problems in graph theory,[1] economic cost/benefit analysis, electronic circuitry design, models of biological evolution, molecular conformational optimization, and protein folding.[2]

A popular method used to isolate the global minimum of a multiextremal objective function is known as ''functional smoothing.'' A simple transformation of the objective function is defined, which produces a smoother function with fewer local minima. When the transformed function has one surviving minimum, that minimum can be found by local minimization methods and mapped back by the inverse transformation to a minimum of the original objective function. Thus, the multiextremal *global* problem is reduced to a series of *local* minimization problems. The hope is that, by this procedure, the global minimum of the objective function is identified.

While the derivative of a function creates an equally or more rugged function, integration renders the function smoother. For example, an integral over the Dirac δ-function gives one, while the derivative of the ramp function is the Heaviside step function. Consider the integral transform of $U(\mathbf{r})$ by a kernel $W_\epsilon(\mathbf{r}, \mathbf{r}_0)$ in the form:

$$\mathscr{U}_\epsilon(\mathbf{r}_0) = \int U(\mathbf{r}) W_\epsilon(\mathbf{r}, \mathbf{r}_0)\, d\mathbf{r} \tag{1}$$

where $U(\mathbf{r})$ is the objective function and $W_\epsilon(\mathbf{r}, \mathbf{r}_0)$ is a window function centered at $\mathbf{r}_0$ with characteristic width ϵ. When ϵ is increased to a length comparable to the separation between minima, the transformation will act to reduce the number of local minima.[5] When ϵ is on the order of the separation between the most distant basins of the objective function, the transformed function may have only a single surviving minimum.

A variety of successful optimization methods have been based on this integral transformation where the smoothing kernel is a Gaussian function[6, 14–16]:

$$W_\epsilon(\mathbf{r}, \mathbf{r}_0) = (2\pi\epsilon^2)^{-d/2} \exp\left(-(\mathbf{r} - \mathbf{r}_0)^2 / 2\epsilon^2\right) \tag{2}$$

where d is the dimension of the space.

For many objective functions, integral transforms using Gaussian kernels are not exactly soluble. In those cases, the objective function may be approximated by a series of polynomial, exponential, or Gaussian functions, which permits the evaluation of eq. (1). Alternatively, numerical quadrature may be used. In either case, the evaluation of $\mathscr{U}_\epsilon(\mathbf{r}_0)$ may be significantly more costly than the evaluation of $U(\mathbf{r})$.

Method of Bad Derivatives

An alternative approach is to approximate the Gaussian kernel by a ''top hat''[7] or impulse function. Although not a standard textbook example,[3] this substitution leads to a very simple evaluation of the gradient of $\mathscr{U}_\epsilon(\mathbf{r})$.

Consider a one-dimensional objective function, $U(x)$, which need not be smooth or differentiable (the function must be integrable). We define the impulse function centered at x_0, and having width 2ϵ, as:

$$I_\epsilon(x, x_0) = \frac{1}{2\epsilon}[\Theta(x - (x_0 - \epsilon)) - \Theta(x - (x_0 + \epsilon))] \tag{3}$$

where the Heaviside function, $\Theta(x)$, is zero for $x < 0$ and unity for $x \geq 0$. Taking $W_\epsilon(x, x_0) = I_\epsilon(x, x_0)$ to be the smoothing kernel, the transformed function is:

$$\mathscr{U}_\epsilon(x_0) = \int_{-\infty}^{\infty} U(x) I_\epsilon(x, x_0)\, dx = \frac{1}{2\epsilon} \int_{x_0 - \epsilon}^{x_0 + \epsilon} U(x)\, dx \tag{4}$$

The smoothing transformation replaces the value of the objective function at x_0 with the average value of the objective function taken over the window of width 2ϵ centered at x_0.[8] The resulting $\mathscr{U}_\epsilon(x_0)$ is smoother than the objective function $U(x)$.[4]

The force derived from the gradient of the smoothed objective function is remarkably simple:

$$\mathscr{F}_\epsilon(x_0) \equiv -\frac{d\mathscr{U}_\epsilon}{dx_0} = -\frac{1}{2\epsilon}[U(x_0 + \epsilon) - U(x_0 - \epsilon)] \tag{5}$$

To compute the force, $\mathscr{F}_\epsilon(x_0)$, associated with the smoothed function $\mathscr{U}_\epsilon(x_0)$, one must simply look up $U(x_0 + \epsilon)$ and $U(x_0 - \epsilon)$. No explicit evaluation of the smoothed function $\mathscr{U}_\epsilon(x_0)$ is required. Eq. (5) appears to be a three-point (central point plus the two extremities), finite-difference approximation to the gradient of the untransformed objective function. This formula provides the exact derivative of the *transformed* function for any value of ϵ.[17] However, we can think of our transformation of U as being crudely based on the force from a bad finite-difference approximation to the derivative of the function—one where ϵ is large.

Repulsions of atoms close to each other give rise to singularities in the potential energy, $U(\mathbf{r})$. This leads to serious problems in optimization methods based on potential energy smoothing and local energy minimization. To overcome this difficulty, it is common to replace potentials of the type $1/r^{12}$, which diverge at $r = 0$ with exponential or Gaussian core repulsions, or to set them to a finite constant at small distances.

It has been recognized[18] that a more appealing solution to the core singularity problem is to coarse-grain a function proportional to the Boltzmann probability density $\rho(\mathbf{r})$ as:

$$\mathscr{U}_\epsilon(\mathbf{r}_0) = \int d\mathbf{r} \exp(-\beta U(\mathbf{r})) I_\epsilon(\mathbf{r}, \mathbf{r}_0) \qquad (6)$$

Singular core repulsions in $U(\mathrm{r})$ are replaced by regions of low probability density in $\rho(\mathbf{r})$, which is a well-behaved, normalized function.

The functional mapping:

$$M(U(\mathbf{r})) = -e^{-\beta U(\mathbf{r})} \qquad (7)$$

of the hypersurface has the effect of reducing the maxima of U and accentuating its minima. In general, for this purpose, any mapping $M(U)$ that is monotonically increasing in U with a first derivative increasing with decreasing U [i.e., $M(U)$ should have negative curvature] would do. The exponential choice of M is reminiscent of the free energy and was chosen as such for its physical significance.

However, a Gaussian integral transform of the probability density is difficult. For all except the most simple potentials, one must resort to numerical quadrature to perform the transform of $\rho(\mathbf{r})$. In contrast, our method is perfectly suited for coarse-graining $\rho(\mathbf{r})$, because only the exponentials of $\beta U(\mathbf{r})$ need to be calculated in the necessary points of the ''bad derivative.''

ITERATIVE MINIMIZATION AND SIMULATED ANNEALING

This ''method of bad derivatives'' (MBD) may be applied for *global* energy minimization using iterative *local* energy minimization on a decreasingly coarse-grained objective function. Initially, ϵ is set to a relatively large value. Subsequently, the following iterative procedure can be performed:

1. Use the force in eq. (5) to perform a steepest descent minimization:

$$\frac{dx_0}{ds} = \mathscr{F}_\epsilon(x_0) \qquad (8)$$

   until a local energy minimum is located.
2. Reduce ϵ by an amount $\Delta\epsilon$.
3. As long as ϵ is greater than zero, go to step 1; otherwise, stop.

The resulting solution gives the best guess for the position of the global energy minimum.

Alternatively, the steepest descent algorithm may be replaced by a simulated annealing procedure employing molecular dynamics, with ϵ playing the role of the ''temperature'' using the following protocol.

Initially, ϵ is set to a relatively large value. Subsequently, follow the procedure:

1. Use the force in eq. (5) to integrate the equation of motion:

$$\frac{d^2\mathbf{r}_0}{dt^2} = \mathscr{F}_\epsilon(\mathbf{r}_0) \qquad (9)$$

   for some period of time to allow for exploration of the smoothed potential surface.
2. Reduce ϵ by an amount $\Delta\epsilon$.
3. As long as ϵ is greater than zero, go to step 1; otherwise, stop.

Another implementation of the method that could improve its efficiency is the following. As explained previously, the bad derivative is a three-point function. Instead of that, one could use a $(2n + 1)$-point scheme, which is equivalent to considering as the smoothing kernel a sum of impulse functions having different heights and different widths.

An important parameter is the ratio $\epsilon/\delta s$, which is a measure of, in computational parlance, the ''stiffness'' of the problem. If the stiffness is too

small, minima can be missed. A lower bound on ϵ can be prescribed and, if the size step is small enough, only narrow minima can be missed; these narrow minima do not make a large contribution to the free energy, as the ''volume'' they encompass is small and contains few configurations.

In contrast to methods based on Gaussian integral transforms, there are no approximations, transforms, or derivatives involved in the computation of the force $\mathscr{F}_\epsilon(x_0)$. Because the method is ''derivative free,'' it is of interest for problems in which one seeks the extremum of a nondifferentiable or stochastic function.[10] As we show in what follows, generalization to multidimensional energy surfaces is straightforward.

MULTIDIMENSIONAL CASE

In this subsection, we discuss the application of our method to the global minimization of a potential function of interacting particles. In particular, we focus on clusters of atoms with nonbonded interactions that can be described by a Lennard–Jones (LJ) potential, and a small peptide, with energetics modeled by an empirical potential with bonded (bond, angle, and torsion) and nonbonded (van der Waals and Coulomb) terms.

To apply the MBD to the conformational optimization of atomic systems in three dimensions, it is necessary to generalize the smoothing kernel. For the particular case of a cluster of atoms, by replacing the one-dimensional top-hat function with a ball of radius ϵ in the three-dimensional space, the centrosymmetric form of the pairwise potential is preserved. The integration kernel W in eq. (1) of the integral transform of the potential function would consist of a product of uniparticle spherically symmetric kernels, having a constant nonzero value inside and a zero value everywhere outside. Each of the uniparticle kernels is centered on a particle. This leads to an analytic form for $\mathscr{U}_\epsilon(\mathbf{r}_0)$ that is rather complicated.[11] Therefore, we consider the generalization of the top-hat function in Cartesian coordinates:

$$I_\epsilon(\mathbf{r}, \mathbf{r}_0) = \frac{1}{(2\epsilon)^d} \prod_{i=1}^{d} \left[\Theta\left(x^i - (x_0^i - \epsilon) \right) - \Theta\left(x^i - (x_0^i + \epsilon) \right) \right] \tag{10}$$

where d is the number of independent variables upon which the potential function depends.

This kernel defines a smoothing of the objective function over the hypercube, $\mathscr{HC}$, centered at $\mathbf{r}_0$, having sides of length 2ϵ parallel to the coordinate axes. Alternatively, it can be looked at as a product of cubic uniparticle kernels, each centered on a particle. The integral transform of the objective function can be written as:

$$\mathscr{U}_\epsilon(\mathbf{r}_0) = \int U(\mathbf{r}) I_\epsilon(\mathbf{r}, \mathbf{r}_0)\, dr = \frac{1}{(2\epsilon)^d} \int_{\mathscr{HC}} U(\mathbf{r})\, d\mathbf{r} \tag{11}$$

$U(\mathbf{r})$ is thus coarse-grained over the hypercube. As in the one-dimensional case, an estimation of the integral in eq. (11) is not needed for global optimization. When the force derived from $\mathscr{U}_\epsilon$ is computed, a major simplification of the formula is obtained. The ith component of the force is:

$$\mathscr{F}_\epsilon^i(\mathbf{r}_0) \equiv -\frac{\partial}{\partial x_0^i} \mathscr{U}_\epsilon(\mathbf{r}_0) = -\frac{1}{(2\epsilon)^d} \int U(\mathbf{r}) \frac{\partial}{\partial x_0^i} I(\mathbf{r}, \mathbf{r}_0)\, d\mathbf{r} \tag{12}$$

$$\cong -\frac{1}{2\epsilon} \left[U(\mathbf{r}_0 + \epsilon \hat{x}^i) - U(\mathbf{r}_0 - \epsilon \hat{x}^i) \right] \tag{13}$$

In performing the integration in eq. (13) it is assumed that the value of the integral of U over a $(d-1)$-dimensional hypersurface of the hypercube equals the value of the function in the center of that manifold. The approximate sign in eq. (13) *can* be replaced by the equal sign if one considers the integral as being performed over a slightly distorted hypercube, or over a variably weighted hypercubic kernel. At any rate, as $\epsilon \to 0$, the exact derivative component is recovered.

By the mean value theorem, the difference between the values of a function at two points is equal to the differential at some intermediate point on the line-segment joining the two points. Should $U(\mathbf{r})$ be twice differentiable, the fact that a minimum of $U(\mathbf{r})$ does not exist in the hypercube implies that $\mathscr{U}_\epsilon(\mathbf{r}_0)$ will not be a minimum either. That is, the smoothing operation will not introduce minima in $\mathscr{U}_\epsilon$ in a region where U has no minima.

The dimensionality of the conformational search space is $d = 3N - 6$, where N is the number of atoms in the system. We eliminate the rigid-body (rotational and translational) degrees of freedom from the potential function. Zero-frequency (null curvature) modes are to be eliminated because they create a flat plateau in the potential energy surface on which the computational search can inefficiently wander.

INITIAL CONDITIONS

How large should we initially set the side ϵ of the hypercube to be such that the smoothed hypersurface has just one minimum? It can be shown by integration with the top-hat kernel that two infinitely narrow minima (modeled by two negative δ-functions following Kostrowicki and Scheraga[19]) separated by distance a merge if $\epsilon = a/2$. This provides an upper bound on ϵ because less deep minima will, in general, join ''faster'' (i.e., for a smaller value of ϵ).

The starting value of ϵ should be of the order of the greatest length in the configuration manifold. This is so, because, to prevent the worst-case scenario, the side of the hypercube should be at least equal to the greatest distance between two points in the accessible region of the multidimensional space. Following the recipe of Kostrowicki and Scheraga, one finds the Euclidean distance, $d(\mathbf{r}, \mathbf{s})$, between the farthest separated configurations $\mathbf{r}$ and $\mathbf{s}$ of a peptide chain when the two chains run in opposite directions. The upper bound for ϵ would be:

$$\epsilon_{max} = d(\mathbf{r}, \mathbf{s}) = \left[\sum_{k=1}^{n} (2lk)^2 \right]^{1/2} = l \left[\frac{n(n+1)(2n+1)}{3} \right]^{1/2} \quad (14)$$

where n is the number of bonds in the backbone of the peptide (the one containing all the amide bonds) and l is a measure of the bond length.

For the Lennard–Jones (LJ) cluster with N atoms there will be, in general, $N!$ points in phase space having the lowest energy minimum. The biggest separation between a pair of them would result from a 180° rotation of the cluster, in which case an upper bound of ϵ would scale as $DN^{1/2}$, where N is the number of atoms and D is the ''diameter'' of the cluster. The same argument holds for the exponential mapping $M(U(\mathbf{r})) = -e^{-\beta U(\mathbf{r})}$ of the potential because the position of the minimum of the mapped potential is the same for the untransformed potential.

SMOOTHING AND SYMMETRY BREAKING

The interactions in atomic clusters are described by central forces. If preserving the centrosymmetry is a major issue, rotations of the cube and averaging should provide stellated polyhedral-shaped kernels better approximating a sphere, thus reducing the effects of the breaking of the centrosymmetry of the problem at hand. Such a polyhedral kernel can be constructed by three coordinate rotations of 45°, around x, y, and z. The three resulting cubes are overlapped onto the initial cube and the broken centrosymmetry is partially healed. More formally, by applying the rotation matrix, $\mathbf{A}$, we get a new vector, $\mathbf{r}' = \mathbf{A}\mathbf{r}$. In this new reference system the forces are $\mathcal{F}_\epsilon(\mathbf{r}')$. These forces are vectors and thus transform just like the coordinates. We have $\mathcal{F}_\epsilon(\mathbf{r}) = \mathbf{A}^{-1}\mathcal{F}_\epsilon(\mathbf{r}') = \mathbf{A}^{-1}\mathcal{F}_\epsilon(\mathbf{A}\mathbf{r})$. Because $\mathbf{A}$ is an orthogonal matrix, $\mathbf{A}^{-1}$ is just $\mathbf{A}^{\mathbf{T}}$. $\mathbf{A}$ is a tridiagonal matrix with 3×3 block matrices on the principle diagonal:

$$\mathbf{A} = \begin{pmatrix} a & & & \\ & a & 0 & \\ & 0 & \ddots & \\ & & & a \end{pmatrix} \quad (15)$$

For example, in the case of a rotation around the z-axis, the block matrices all have the form:

$$a = \begin{pmatrix} \sqrt{2}/2 & \sqrt{2}/2 & 0 \\ -\sqrt{2}/2 & \sqrt{2}/2 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad (16)$$

The calculation is repeated for rotations around the other two axes and new forces are computed after which the average force of the four forces (the force in the original reference frame plus the three rotated ones) is evaluated and used as the effective force in the MBD algorithm. Such an approach brings extra computational requirements (a fourfold increase in the number of potential energy function calls) with modest improvement in the approximate conservation of the centrosymmetry of the system and little effect on the quality of the isolation of minima.

Interpretation

The smoothing transform is a convolution integral, so it is natural to analyze the effect of smoothing in Fourier space. Because $\mathcal{U}_\epsilon(x)$ is the convolution of $U(x)$ and $W_\epsilon(x)$, by the convolution theorem we have:

$$\tilde{\mathcal{U}}_\epsilon(k) = \tilde{U}(k)\tilde{W}_\epsilon(k) \quad (17)$$

where $\tilde{\mathscr{U}}(k)$ is the Fourier transform of the function $U(x)$. If the kernel is Gaussian with standard deviation ϵ as in eq. (2), then:

$$\tilde{W}_\epsilon(k) = \exp(-\epsilon^2 k^2/2) \tag{18}$$

and the effect of the Gaussian smoothing is clear. The fluctuations in the potential energy with a length scale $1/k < \epsilon$ are filtered out of the smoothed potential function.

The case is more complicated for the top-hat smoothing function. The Fourier transform of the smoothing function of width 2ϵ is:

$$\tilde{W}_\epsilon(k) = \frac{1}{\epsilon k}\sin(\epsilon k) \tag{19}$$

recognizable as a single slit diffraction pattern amplitude. This smoothing transform acts to remove the longer length scale variations in the objective function as does the Gaussian smoothing. The attenuation is more gradual than in the case of a Gaussian smoothing. Moreover, the smoothing is not monotonic. There are oscillations and periodic zeros in $\tilde{\mathscr{W}}_\epsilon(k)$. This feature complicates the general analysis of the effect of the top-hat smoothing in Fourier space.

Let us consider the more specific case of the function:

$$U(x) = U_0(x) + \Delta U(x) \tag{20}$$

which consists of a background term $U_0(x)$, a smoothly varying function of x, and a term $\Delta U(x)$ that adds ruggedness. If we decompose the ruggedness in a Fourier series:

$$\Delta U(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty}[a_n\cos(nx) + b_n\sin(nx)] \tag{21}$$

If the smoothing kernel is a top-hat function, as in eq. (3), the smoothed ruggedness will have the form:

$$\Delta\mathscr{U}_\epsilon(x_0) = \frac{a_0}{2} + \sum_{n=1}^{\infty}\frac{1}{\epsilon n} \times \sin(\epsilon n)[a_n\cos(nx_0) + b_n\sin(nx_0)] \tag{22}$$

When the smoothing length scale $\epsilon = N\pi/n$, where N is an integer, the sinusoidal ruggedness of wavelength $L/n\pi$ will entirely disappear. As ϵ is increased beyond that point, the ruggedness will reemerge, but it will be attenuated.

Consider the one-dimensional potential, $U(x)$, where the smoothing function is centered at $x_0 = 0$. For the Gaussian smoothing kernel, the result is:

$$\mathscr{U}_\epsilon(x_0) = U(0) + \sum_{n=1}^{\infty}(2n-1)!!\epsilon^{2n}\frac{1}{(2n)!}U^{(2n)}(0) \tag{23}$$

For the top-hat smoothing kernel, the result is:

$$\mathscr{U}_\epsilon(x_0) = U(0) + \sum_{n=1}^{\infty}\frac{1}{2n+1}\epsilon^{2n}\frac{1}{(2n)!}U^{(2n)}(0) \tag{24}$$

The higher order derivatives of $U(x)$ enter with a much smaller proportionality in the case of the top-hat smoothing function. This is due to the more delocalized nature of the Gaussian kernel. Unlike the top-hat function, the Gaussian kernel's ''feet'' extend far beyond its ''waist.''

This feature is most apparent for the case of the quartic double-well potential of the form:

$$U(x) = (x^2 - 1)^2 \tag{25}$$

which has minima at $x = \pm 1$ and a barrier at $x = 0$ of height unity. The potential smoothed over a top-hat kernel is:

$$\mathscr{U}_\epsilon(x_0) = x_0^4 - 2x_0^2(1-\epsilon^2) + (1 - 2\epsilon^2/3 + \epsilon^4/5) \tag{26}$$

which has two minima for $\epsilon < 1$ positioned at $x_0 = \pm\sqrt{1-\epsilon^2}$; for $\epsilon \geq 1$ the smoothed potential has a single minimum at $x_0 = 0$. The smoothed potential computed using a Gaussian kernel is:

$$\mathscr{U}_\epsilon(x_0) = x_0^4 - 2x_0^2(1-6\epsilon^2) + (1 - 2\epsilon^2 + 3\epsilon^4) \tag{27}$$

which has two minima for $\epsilon < 1/6$ positioned at $x_0 = \pm\sqrt{1-6\epsilon^2}$; for $\epsilon \geq 1/6$, the smoothed potential has a single minimum at $x_0 = 0$.

What conclusions can we draw? Top-hat smoothing leads to a softening of the walls and barriers of the potential (with higher order moments being attenuated). Gaussian smoothing rapidly annihilates barriers, mostly through thickening of the walls. Length scale ϵ is an exact measure of the smoothing length in the MBD transform; minima separated by ϵ or less are

joined. For Gaussian smoothing, ϵ is a length significantly shorter than the effective smoothing length.

Applications

In this section we apply the MBD to a model potential and then use it to perform conformational optimization for a series of atomic clusters and a tetrapeptide.

ONE-DIMENSIONAL MODEL POTENTIAL

To illustrate the one-dimensional method, we choose an objective function of the form:

$$U(x) = x^2 - ae^{-b(x+2)^2} - ce^{-d(x-2)^2} - fe^{-g(x-3)^2} \quad (28)$$

consisting of a quadratic background with three Gaussian minima drilled in. It is illustrated in Figure 1 for $a = f = 15$, $b = c = 10$, and $d = g = 3$, along with successive transformations using the top-hat kernel. Figure 2 shows the successive transformations of the same function, $U(x)$, when Gaussian smoothing is applied. For the choice of parameters in eq. (28), the MBD finds the correct minimum. The Gaussian method finds a local minimum in the wider basin of attraction (catchment region). It could be argued that this difference might be a feature related to the particular choice of parameters. However, a careful study shows that the range of parameters describing the shape of the function in eq. (28) for which the MBD finds the global minimum contains, as a subset, the range of parameters for which the Gaussian smoothing is successful. The varying success can be explained as follows. Consider first the integral in eq. (1) performed using the Gaussian kernel. The contribution of regions with highly repulsive walls to the integral is nonzero. In fact, the smoothed function will diverge if the walls rise faster than $\exp(x^2)$. When low-lying local minima are surrounded by steeply repulsive walls, the value of the integral can be very high, filling in (annihilating) that cluster of minima. This does not happen to such an extent when the top-hat kernel is employed. This is the effect mentioned in the previous section and in the discussion of the quartic double-well potential. Similar results were obtained for the Boltzmann probability of eq. (6) as applied to our one-dimensional model potential (see Figs. 3 and 4).

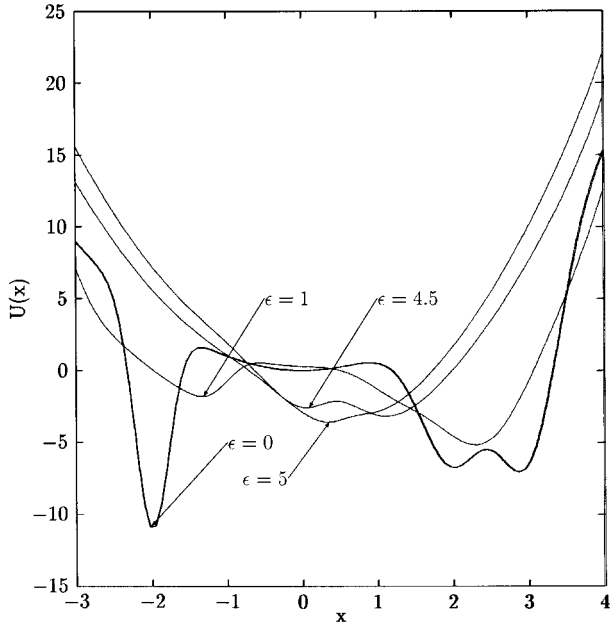

FIGURE 1. The method of bad derivatives is used to smooth the potential function, eq. (28). The transformed potential is overlapped on the untransformed potential for different values of parameter ϵ. The arrows also indicate the position of the minimum traced by local minimization for each of the curves. By tracing back the position of the minimum the method of bad derivatives succeeds in locating the global minimum of the original potential.

ATOMIC CLUSTERS

Using the multidimensional smoothing scheme based on the kernel in eq. (10), we performed searches for the global potential energy minimum of a series of Lennard–Jones clusters. We first estimated the size, ϵ, of the hypercube side, which we subsequently decreased exponentially. Initially, atoms were randomly placed in a cube. We performed minimization by the conjugate gradient method using derivatives on the smoothed potential surfaces. The CPU timing was, in all the cases we studied on the order of 10^3 seconds on a Silicon Graphics 150-MHz Challenge/S server. No special steps were taken to optimize the application program.

The potential of the cluster is pairwise additive, the interaction of each pair of atoms being modeled by a Lennard-Jones potential expressed in

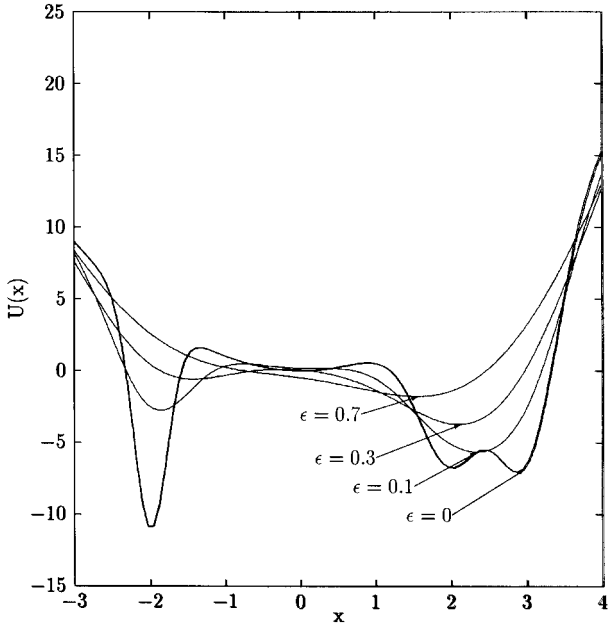

FIGURE 2. The potential function, eq. (28), is transformed using the Gaussian kernel. We show the transformed potentials for different values of the Gaussian width overlapped on the untransformed potential. The method fails by locating a local energy minimum.

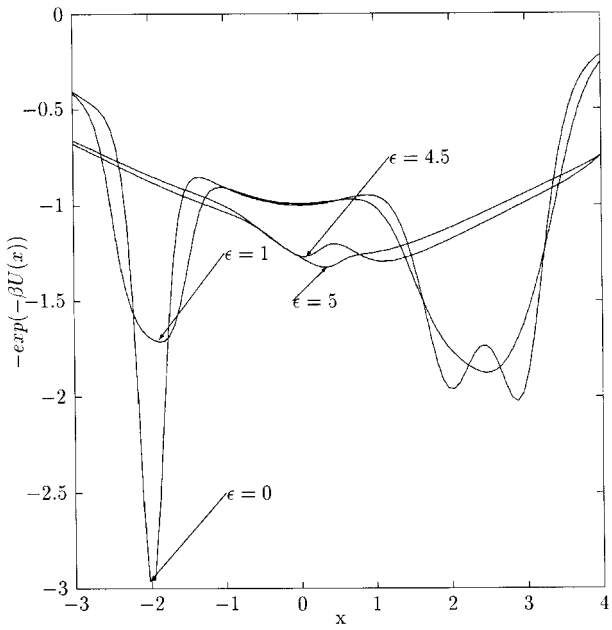

FIGURE 3. The method of bad derivatives applied to find the minimum of $M(U) = -\exp(-\beta U(x))$ using $\beta = 1/10$ in order to find the global minimum of the potential function, eq. (28). The position of the global minimum is found correctly.

energy units of the well depth and length units of the pair equilibrium distance:

$$u(r) = \frac{1}{r^{12}} - \frac{2}{r^6} \tag{29}$$

where r is the interatomic distance. To confine the atoms during the conjugate-gradient search, we also add to the Lennard–Jones pairwise interaction a confining boundary potential of the type $(r/r_0)^{20}$, where r_0 is greater than the size of each cluster. The confining potential has little effect on the small interatomic distances. However, it does control the direction in which the minimum of the pair potential shifts. As $r \to 0$, if the confining boundary potential rises, as r goes to infinity more slowly than the interatomic repulsion potential rises, then the potential minimum will be shifted toward larger interatomic distances.

The elimination of the six rigid-body degrees of freedom, discussed in the ''Method of Bad Derivatives'' section, is done by fixing the first atom of the cluster to lie at the origin of the xyz Cartesian reference frame, the second along the Ox axis, and the third atom in the xOy plane. We searched for

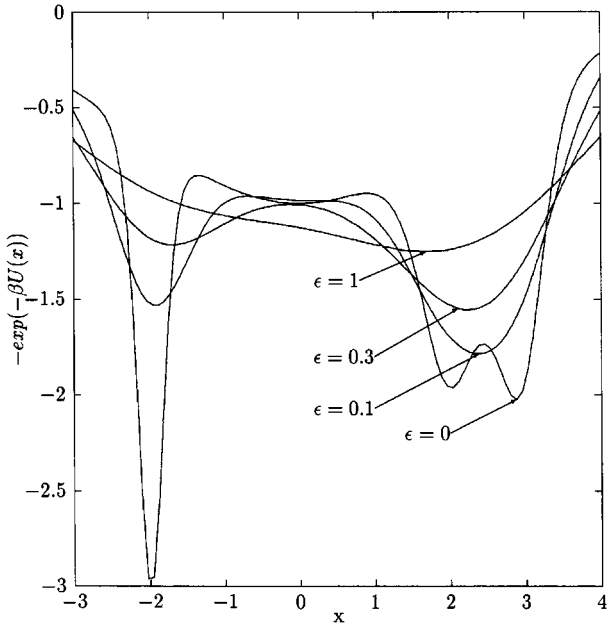

FIGURE 4. Gaussian smoothing is applied to find the minimum of $M(U) = -\exp(-\beta U(x))$ using $\beta = 1/10$ in order to find the global minimum of the potential function, eq. (28). The method fails by locating a local energy minimum.

the global minimum (known from previous growth studies[12]) for clusters with numbers of atoms ranging from 5 (the first nontrivial case, two minima) up to 16. In the numerical implementation of the algorithm we found that the use of the untransformed potential energy function is more computationally unstable than a smoothing of the mapping of the potential function, using exponential mapping in eq. (7). The reason is that the minimization method uses the gradient that includes a large contribution from atom–atom repulsions giving rise to divergencies when an edge of the smoothing kernel is close to such a repelling atom.

Table I shows the minima found by smoothing and iterative minimization of the smoothed hypersurface:

$$\mathscr{U}_\epsilon(\mathbf{r}_0) = -\int e^{-\beta U(\mathbf{r})} I_\epsilon(\mathbf{r}, \mathbf{r}_0)\, d\mathbf{r} \qquad (30)$$

with $\beta = 1/20$ for the clusters up to 10 and $\beta = 1/40$ for the remainder of the clusters. The units of energy are reduced Lennard–Jones units. From the nine-atom cluster on, the name of the packing geometry that includes the global minimum is provided. In parentheses are the known global minima derived from the thorough studies of Hoare and Pal.[12]

ALL-ATOM MODEL PEPTIDE

We applied the MBD algorithm to the conformational optimization of the peptide [isobutyl-$(ala)_3$-methlyamide], using, as the function to be smoothed, an empirical model of the potential energy of the molecule and the coarse-grained probability density in eq. (6). For this empirical model we employed the CHARMM force field,[13] containing harmonic bond and angle terms and dihedral, electrostatic, and Lennard–Jones terms. There was no potential energy truncation.

TABLE I.
MBD Applied to Maximization of Boltzmann Probability for Lennard – Jones Atomic Clusters.

Atoms	Found (global)	Configuration
5	−9.104 (−9.104)	Trigonal bipyramid
6	−12.712 (−12.712)	Octahedron
7	−16.505 (−16.505)	Pentagonal bipyramid
8	−19.819 (−19.819)	Pentagonal bipyramid +1 atom
9	−24.112 (−24.112)	Pentagonal bipyramid +2 adjacent
10	−28.420 (−28.420)	Pentagonal growth
11	−31.945 (−32.765)	Tetrahedreal (pentagonal)
12	−37.967 (−37.967)	Pentagonal growth
13	−44.327 (−44.327)	Pentagonal growth
14	−47.845 (−47.845)	Pentagonal growth
15	−52.322 (−52.322)	Pentagonal growth
16	−56.815 (−56.815)	Pentagonal growth

This peptide's configuration of the global minimum is known from a study involving an extensive search of the configurational space, done previously by Czerminski and Elber, who mapped most of the energy landscape.[20] The starting configurations were obtained from a molecular dynamics trajectory at 3000 K. The integration of the dynamical equations of motion was performed using the velocity–Verlet integrator.[21] As in the case of the clusters, 6 degrees of freedom were eliminated from three atoms. The first atom was "pinned" to the origin of the coordinate system, the second allowed to move on a coordinate plane, and the third on a coordinate axis.

After several different initial configurations of the peptide were generated, conjugate-gradient descent was performed on the smoothed potential and also on the smoothed exponential mapping of the potential. The initial value of the side of the smoothing hypercube was chosen, based on eq. (14), to be $\epsilon = 150$ Å. The initial (large ϵ) minimum energy structures obtained from multiple initial configurations were overlapped to be the same. This indicates that ϵ was large enough to yield one minimum (assuming that the generated initial configurations spanned most of the available configurational space).

The MBD based on the smoothing of the Boltzmann probability [eq. (6)] yielded the global minimum of the potential energy function. The MBD algorithm applied to $U(\mathbf{r})$ itself (with the same schedule of decreasing the value of ϵ) resulted in a local minimum. Of course, maximization of the Boltzmann probability is equivalent to minimization of the potential energy. However, it is better to apply the MBD to the maximization of the Boltzmann probability rather than the direct minimization of the potential energy itself.

Figure 5 represents the peptide in a ball-and-stick model in the configurations isolated by potential smoothing and "free-energy" smoothing, both according to the MBD algorithm. The overlap of the structure obtained by the MBD algorithm using potential smoothing on the global minimum structure was only 2.3 Å, and the difference be-

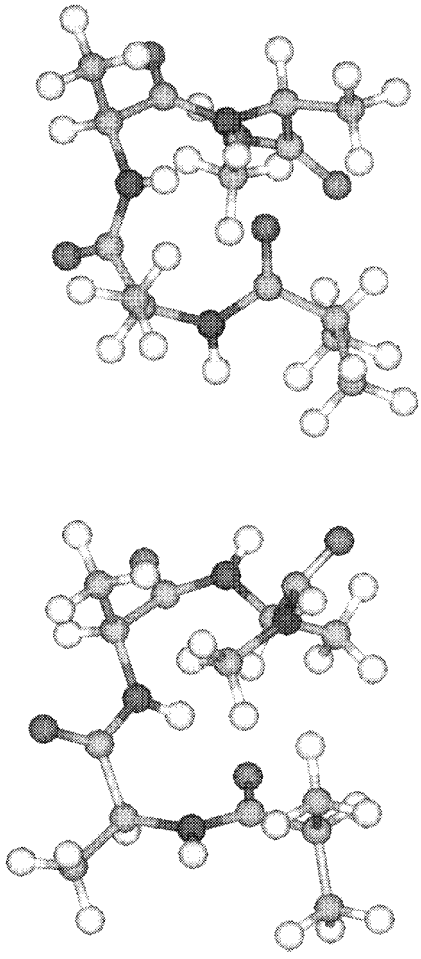

FIGURE 5. Ball-and-stick representations of the peptide tetraalanine in the conformations obtained by smoothing of the functions $M(U) = -\exp(-\beta U(\mathbf{r}))$ (up) and $U(\mathbf{r})$ (down).

tween the values of the potential energy was about 9 kcal/mol. The global minimum structure had three hydrogen bonds, whereas the local minimum structure had one hydrogen bond.

Conclusions

We have presented the MBD, a general method for conformational optimization based on a coarse-graining of the objective function. The algorithm is easy to implement and needs only the value of the objective function. We demonstrated its efficiency by applying it to one-dimensional as well as multi-dimensional examples.

In the limit of a continuous decrease of parameter ϵ, the MBD is deterministic in the sense that it can track back the position of a minimum found at the largest value of ϵ (when there is presumably only one minimum). However, computer implementations involve discretizations and *a priori* knowledge of how slow the reversal of the smoothing must be is unknown. Each smoothing method can fail to isolate the global minimum if two nearly degenerate minima are close in configuration space. The analysis presented in the ''Interpretation'' section, describing the spreading of the Gaussian kernel as opposed to the cubic kernel, demonstrates that the MBD is less susceptible to failure than Gaussian smoothing methods.

Acknowledgments

The authors warmly thank Francesca Massi for technical assistance and Claudio Rebbi for helpful comments.

References

1. E. Aarts and J. Korst, *Simulated Annealing and Boltzmann Machines*, Wiley, New York, 1989.
2. P. M. Pardalos, D. Shalloway, and G. Xue, Eds. *Global Minimization of Nonconvex Energy Functions: Molecular Conformation and Protein Folding*, American Mathematical Society, Providence, RI, 1996.
3. S. Redner (personal communication): incidentally, a similar approximation gives the correct long-time survival probability of the random walk in first passage time problems.
4. Scheraga has argued that, for a Gaussian smoothing transformation, the catchment region with the greatest area will be the last surviving minimum of the transformed function. This is true in one dimension when the local minima of the objective function are well separated. In that case, the magnitude of the integral over each well is most negative for the well of greatest volume. For the Gaussian kernel the separation between minima that is required for this argument to hold will be greater than for the case of the ''top-hat'' smoothing function. It follows that the argument that the last surviving minimum in the smoothing transformation will be the minimum of greatest volume is stronger for the case of the method of bad derivatives.
5. V. V. Zakharov, *Eng. Cybernet.*, **4**, 637 (1970).
6. L. Piela, J. Kostrowicki, and H. A. Scheraga, *J. Phys. Chem.*, **93**, 3339 (1989); J. Kostrowicki, L. Piela, B. J. Cherayil, and H. Scheraga, *J. Phys. Chem.*, **95**, 4113 (1991).
7. S. F. Edwards and D. Wilkinson, *Proc. R. Soc.*, **A381**, 17 (1982).

8. The method is related to methods of filtering phase-space trajectories of chaotic signals for the purpose of eliminating stochastic noise (as opposed to deterministic noise). In a typical example,[9] the phase-space trajectory of a noise-contaminated evolution is smoothed by replacing the geometric center of a suitably chosen neighborhood by its mass center.
9. T. Schreiber, *Phys. Rev.*, **E47**, 2401 (1993).
10. J. Kiefer and J. Wolfowitz, *Ann. Math. Stat.*, **23**, 462 (1952).
11. I. Andricioaei and J. E. Straub (unpublished results).
12. M. R. Hoare and P. Pal, *Adv. Phys.*, **20**, 161 (1971).
13. B. R. Brooks, R. Bruccoleri, B. Olafson, D. States, S. Swaminathan, and M. Karplus, *J. Comput. Chem.*, **4**, 187 (1983).
14. P. Amara, D. Hsu, and J. E. Straub, *J. Phys. Chem.*, **97**, 6715 (1993); J. Ma, D. Hsu, and J. E. Straub, *J. Chem. Phys.*, **99**, 4024 (1993); J. Ma and J. E. Straub, *J. Chem. Phys.*, **101**, 533 (1994); *J. Chem. Phys.*, **103**, 9113 (1995); J. E. Straub, J. Ma, and P. Amara, *J. Chem. Phys.*, **103**, 1574 (1995); P. Amara and J. E. Straub, *J. Phys. Chem.*, **99**, 14840 (1995); P. Amara and J. E. Straub, *Phys. Rev. B*, **53**, 13857 (1996).
15. J. E. Straub, in *Recent Developments in Theoretical Studies of Proteins*, R. Elber, Ed., World Scientific, Singapore, 1996; P. Amara, J. Ma, and J. E. Straub, in *Global Minimization of Nonconvex Energy Functions: Molecular Conformation and Protein Folding*, P. M. Pardalos, D. Shalloway, and G. Xue, Eds. American Mathematical Society; I. Andricioaei and J. E. Straub, *Comput. Phys.*, **10**, 449 (1996).
16. F. H. Stillinger and T. Weber, *J. Stat. Phys.*, **52**, 1429 (1988); F. H. Stillinger, *Phys. Rev.*, *B***32**, 3134 (1985).

17. Similarly, when a ''tent'' smoothing kernel is employed:

$$W_\epsilon(x; x_0) = T_\epsilon(x; x_0) = \begin{cases} (\epsilon - |x - x_0|)/\epsilon, & \text{if } |x - x_0| \le \epsilon \\ 0 & \text{if } |x - x_0| > \epsilon \end{cases} \quad (31)$$

the expression for the *second* derivative of the smoothed function takes on the simple form:

$$\frac{d^2\mathscr{U}}{dx_0^2} = \int_{-\infty}^{\infty} U(x) \frac{\partial^2}{\partial x_0^2} T_\epsilon(x; x_0)\, dx = \frac{U(x_0 + \epsilon) - 2U(x_0) + U(x_0 - \epsilon)}{\epsilon^2} \quad (32)$$

As for the first derivative, this exact second derivative of the smoothed function takes the form of a three-point finite-difference approximation to the second derivative of $U(x)$.

18. D. Shalloway, in *Recent Advances in Global Optimization*, C. A. Floudas and P. M. Pardalos, Eds., Princeton University Press, Princeton, NJ, 1992, p. 433.
19. J. Kostrowicki and H. Scheraga, *J. Phys. Chem.*, **96**, 7442 (1992).
20. R. Czerminski and R. Elber, *J. Chem. Phys.*, **92**, 5580 (1990).
21. M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids*, Oxford University Press, New York, 1990.